The search for genes associated to disease: The LASSO method and the case of Bone Mineral Density

Henrik B. Roald

April 10, 2015

Abstract

Bone Mineral Density (BMD) tells us about the amount of minerals in bones, and gives an indication of bones strength. The BMD is influenced by genetic factors as well as nutrition, sex hormone status, glucocorticoid therapy and physical activity among others. In this project we have focused on genes making an impact on BMD, by using mathematics and regression. Since the number of explanatory variables x (genes) overrides the number of response variables y (women) we use the variable selection method called LASSO. In that way we are able to sort out the genes with the highest contribution to BMD. We discovered 55 unique genes, among which *ATP5SL, TNXA /// TNXB, SOST, PBXIP1, SNCA, GTF2F2, DPP8, C4orf31, TFDP2 and MCF2L* were among those with the strongest association to BMD. *SOST, AFFX-M27830_M_at, PBXIP1* and *RNF216* found by LASSO, confirms previous literature. Most of the genes seems to be highly associated with and probably have some impact on BMD.

Contents

1 BMD - Bone Mineral Density

Bone Mineral Density (BMD) tells us about the amount of minerals in bones, and gives an indication of bones strength. Geometry, micro-architecture, minerals and extracellular matrix composition makes up a bone's construction and strength. There is an association between high and low BMD and strong and fragile bone, respectively [16]. A reduction of BMD and deterioration of the micro architecture of bone tissue, is a sign of either *ostopenia* or *osteoporosis* (OP), and increase the risk of bone fractures [16]. In general practice, BMD is usually measured by dual-energy X-ray absorptiometry (DEXA). The World Health Organization (WHO) *defines osteoporosis as a BMD* $\leq -2,5$ *standard deviations below the mean value for a young healthy adult. Values between* $-2,5$ *and* -1 *are defined as Osteopenia.* [11]. There is a clear inheritability of characteristics of bone mineral density, and examples of genes found earlier to have an impact on bone, are bone morphogenetic protein 2 (*BMP2*), low-density lioprprpotein receptor related protein 5 (*LRP5*) and osteoprotegrin (*TNFRSF11B*). Other factors known to influence BMD, are nutrition (e.g. Calcium insufficiency), sex hormone status (especially oestrogen deficiency), glucocorticoid therapy and physical activity [17]. It is the most prevalent metabolic bone disorder worldwide. One in three women and one in five men over the age of 50 are affected [12].

In the U.S. and Scandinavia, rates are about 25% higher for getting a fracture due to OP compared to other European countries. In the U.S. 40% of all Caucasian women and 13% of all Caucasians of similar age will suffer a fracture of clinical significance due to OP during their lives. These fractures have consequences on both population healthcare and economical/societal aspects, and is of big importance due to an increasingly ageing population [16].

2 Genomics

Genes hold the information of what kind of product a cell shall produce. Most often this is proteins, but it can also be different kinds of RNA such as *ribosomal RNA* ($rRNA$), which functions as a component in ribosomes (complexes responsible for protein synthesis) and *Transfer RNA* ($tRNA$) [2] carries its specific amino acid to the ribosomes for protein synthesis. *Small nuclear RNA* ($snRNA$) plays a role in modulation of *messenger RNA*($mRNA$) [2] which is the template for the proteins that is to be produced. What kind and amount of product that are produced, depend on the regulation of the *gene expression* of the cell, and is regulated in many different ways at both the transcriptional level, where RNA is transcribed and modulated from DNA, and at the translational level, where the modulated mRNA are translated to proteins [2]. The genetic code stored in a cell's DNA gives the base for the gene expression. In the

code, we find promoters, which are the place on the DNA where the transcription starts. Together with hormones, vitamins and proteins called transcription factors, the first part of the gene expression process is initiated here with an enzyme called *RNA polymerase* [2]. The genes serve as a code dictating how the sequence of a RNA-strand should be put together. In eukaryotic organisms, it is first transcribed a $pre-mRNA$, that is further modificated to become a mature $mRNA$. One of the modulation processes is called *splicing*. Here, interfering segments called *introns* are removed, and the remaining segments called *exons* are put together in new ways. This gives rise to many different phenotypes from one single gene. After undergoing some more changes, the mRNA is further translated to proteins, and regulation may also occur at this level [2]. In addition to regulation at the transcriptional and translational level, the DNA may undergo epigenetic changes, which means that the gene expression is changed without interference of the DNA sequence. This is done by changes to the access to the DNA. The amount of copies of a gene may also make an influence, as well as other processes we will not mention further here. It is widely known that genes are inheritable, and traits seen in an organism are, among other, controlled by the gene expression-process. Interesting for the present work is for example that mothers with lower bone mass, more frequently have children with significantly lower bone mass than the general female population [4].

3 How to analyze gene expression

In order to find out which genes that show the greatest impact on a disease, we need to compare the gene expression in individuals with the disease and without the disease. This is done by detecting and measuring the final gene products, RNA and proteins. There are several techniques to do this and some of the most common are presented here:

Northern blotting gives a measurement of the amount and size of particular mRNA molecules. Through electrophoresis are mRNA molecules separated and then transferred to a membrane where they are hybridized to a radioactive probe of the sequence wanted to be analysed. Band-analysis of the membrane with the radioactive probes are done by autoradiography, and gives a measure of the amount and size of the chosen mRNA[2].

PCR - Polymerase Chain Reaction is another approach for measuring mRNA. A single-stranded DNA or RNA are made from a targeted gene sequence from the cell tested through reverse transcription. The product, *cRNA/cDNA* is then amplified through several cycles of replication. The fluorescence from labelled hybridization probes or intercalating dyes are measured. From this you can get an absolute measurement of the number of copies of original mRNA. This method is very sensitive, and theoretically is it possible to detect a single mRNA

molecule as long as we know the sequence that is needed for initiation of the process[14].

To analyse many genes within a sample simultaneously, *hybridization microarray* is used. This is the method that has been used in this project. A microarray is a chip containing probes to every known gene in the genome of an organism. The principle is to convert mRNA to cDNA labelled with a fluorescent tag. The cDNA is then exposed to the chip. By measuring the amount of fluorescence at each spot, we are able to measure the amount of that particular mRNA and thereby discover if there are any differences in gene expression between a diseased and healthy individual or between two different cells from different locations on the body[2].

Other methods, such as Serial analysis of gene expression (SAGE) and RNA-Seq, can provide relative measure of the cellular concentration of different mRNAs. RNA-Seq can also be used to identify single-nucleotide polymorphisms (SNPs), splice variants and novel genes in addition to profile expression in organisms where little or no sequence information are available[18].

Some of the methods used to analyze proteins, are among other Enzyme-linked immunosorbent assays (ELISA). Here antigen, the protein, is bound to a microtiter dish due to an antibody specific for that antigen at the bottom of the dish. The probes that are used, are antibodies specific for the different proteins and they are covalently bound to an enzyme that will change colour of its substrate when exposed to it. The amount of colour from the transformed substrate can tell us about the amount of protein.

Another way is to use *Western Blot*, which is quite similar to Northern blotting. Instead of RNA-molecules, protein molecules are separated by electrophoresis and transferred (blotted) to a membrane. A antibody labeled probe is also used here, and it produces a band at the location of its antigen on the membrane that is further analyzed [2].

4 Bones and BMD

4.1 Materials

Mature bone consists of a mixture of about 70% inorganic salts and 30% organic matrix. Of the organic matrix, 90% is *collagen* and the rest being *ground substance proteoglycans* and a group of *non-collagen molecules* involved in the regulation of bone mineralisation. The ground substance *proteoglycans*' task is among other controlling the water content of bones and probably regulating formation of collagen fibers in order to get an appropriate subsequent matrix mineralisation. Other non-collagen organic material is *osteocalcin*, responsi-

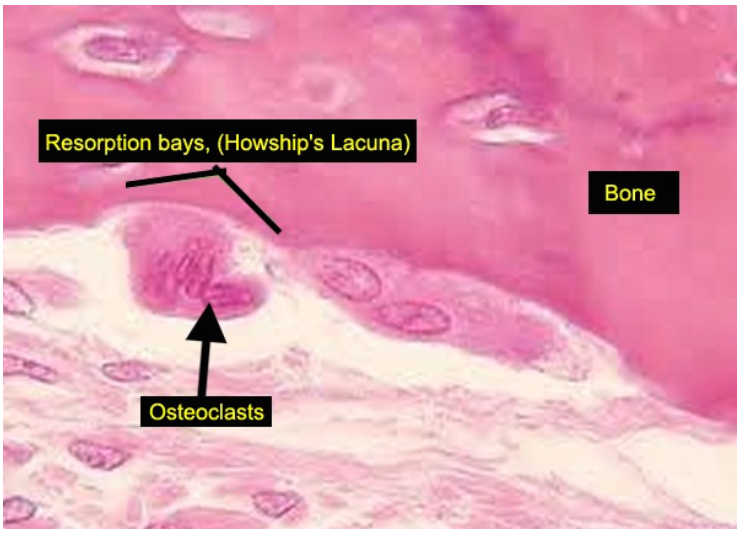


Figure 1: Howship lacunae with osteoclast[9].

ble for binding Calcium during the mineralisation process, *osteonectin*, which probably serves some bridging function between collagen and the mineral component, *sialoproteins*, a surface glycoprotein and certain other proteins [19]. The inorganic component consist of calcium and phosphate crystallized as hydroxyapatite, and is conjugated with small portions of magnesium carbonate, sodium and potassium [19].

4.2 Osteoblasts and osteoclasts

Anabolic cells called *osteoblasts*, have the responsibility of synthesis and secretion of *collagen* and other organic matrix which results in the production of osteoid. They also have the responsibility of deposition of calcium salts in the bone matrix [19]. As the osteoblasts produce organic matrix, they get surrounded by this matrix and eventually get trapped and transformed to *osteocytes*. The Osteocytes' main responsibility is to take part in the mineral homeostasis of the bone matrix. The cell body of the osteocyte makes a space called *lacuna*, and the cells dendrites make channels in the matrix called *canaliculi*. Through these canaliculies, the cell gets its nutritients. In the bone there are also katabolic cells called *osteoclasts*. They resorb bone through the enzyme *acid phosphatase* and leave behind cavities called *Howship lacunas*, see **Figure 1**. One osteclast can resorb as much bone matrix as 150 osteoblasts can produce per unit of time [6].

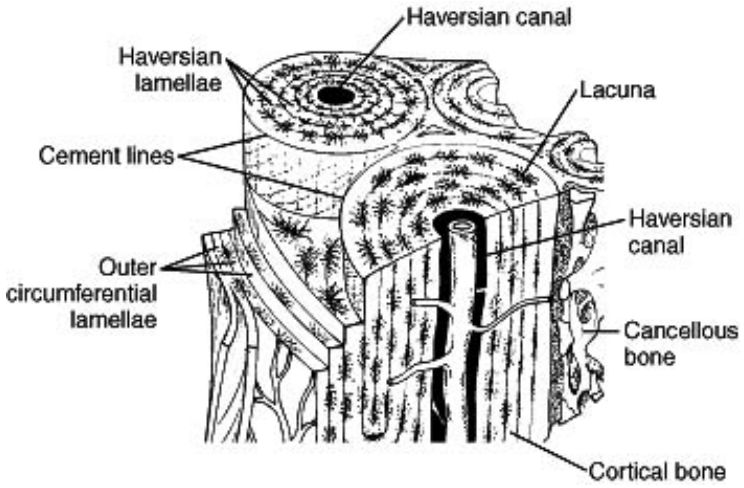


Figure 2: Howersian canal with surrounding matrix [3].

4.3 Anatomy

In the microscopic bone matrix, there are small units called *Osteons* or *Haversian systems*, see **Figure 2**. These units are made up of concentric shells or bone lamellae, *Haversian lamellae*, and they form a channel called *Haver's channel*. This channel contains capillares and nerves. The *Haversian systems* are made by osteoblasts which produce layers of lamellae from the center of the osteon, and eventually ends up being trapped between the lamellaes as osteocytes. In young individuals, the osteons lays paralell, and as we get older the lamellae will overlap each other. Macroscopically, the bones are made up of to types of bones, compact bone and cancellous bone. The compact bone makes up all parts of the skeletal surface. The cancellous bone is found mainly in the end of the bones and in the short bones. It consist of a network of supporting beams, *trabeculae*, which direction are decided by the bones tension and compression lines, *trajectories*. This give the bones the property of having maximum strength with a minimum of material [6]. In osteoporosis a widening of the haversian canals is seen microscopically as well as a macroscopically thinning and widely separation of the trabeculae. It happens due to increased resorbtion by osteoclasts, reduced bone formation by osteoblasts and reduced maintenance of osteocytes [17].

5 Bone and genomics

In order to make some impact on treatment or prevention of OP, several studies have been done regarding the genetic significance for BMD. In an icelandic study on first degree relatives in several families, it has been discovered that few

genes influence low bone mass considerably, in addition to several other genes with a small effect [16].
Some twin studies have showed that bone mass is about 80% genetically determined. Single-nucleotide polymorphisms (SNPs) have also been identified to make an impact on BMD, but through wide genomic scan studies, it is estimated that these account for only about 4% of the total variation in hip and spine BMD [11]. Other smaller studies have tried to describe gene expression related to BMD and OP in bone biopsies. One of them is the Hopwood study, which consisted of 10 bone autopsy samples (controls), 10 biopsies from patients with osteoarthritis (OA, second control) and 10 with OP. The study showed 150 differentially expressed genes in OP-bone compared to OA and control-biopsies [16].
In this project, we will try to identify which genes are associated with BMD variation through regression, using the variable selection-method called LASSO.

6 Linear regression

6.1 Univariate linear regression

The simplest type of regression, is linear univariate regression. It is used when we are studying the relation between an explanatory variable x (for example the expression of a single gene) and a response variable y (for example BMD). In many situations, we do not know this relation, but we can assume that it is linear of the form $y = a+bx$, where a is the intersection point on the Y-axis and b is the slope of the line. If you have n number of points $(e_1, e_2, e_3, ..., e_n)$, each with the value (x_1, y_1), (x_2, y_2), (x_3, y_3),..., (x_n, y_n), our main objective is to do a regression and estimate the values for a and b so that the vertical distance to the line $y = a + bx$ is as small as possible for all the points, see **Figure 3**.

This is normally done by using the least square method, expressed by

$$\min_{a,b} \sum_{i=1}^{n} (y_i - a - bx_i)^2.$$

As seen in the figure, this formula takes the distance from the line to each point, squares the distances in order to get positive values and in the end sums all the distances. The values of a and b are found by deriving the formula, and putting the derivative equality, in order to find the minimal extremal of the graph [1].

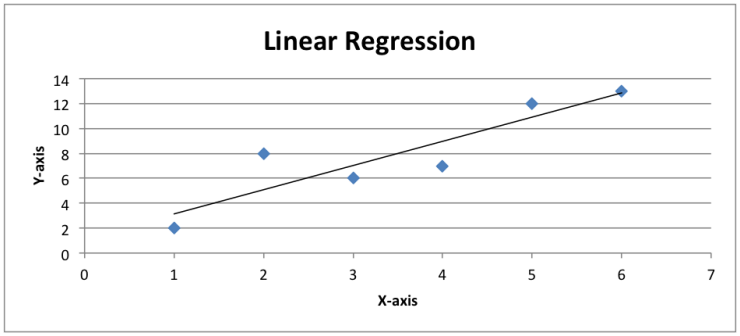


Figure 3: Linear regression example

6.2 Regression with several independent variables

6.2.1 Linear multivariate regression with p<n

In medicine, there is seldom only one variable that makes an impact on the response variable y (we still use BMD as an example for y). Variables such as height, weight, hormones etc. may have an influence on the BMD. This is expressed by $(x_{1i}, x_{2i}, ..., x_{pi}, y_i)$, where p is the number of influencing variables, y is the BMD and i is the index of the different samples, ranging from $(i = 1, ..., n)$. Instead of a line, there is now a function for a plane in a *p+1*-dimensional room with the coordinates y and p numbers of x-coordinates/axes, see **Figure 4**.

Working with more than one explanatory variable x, the regression formula will be expressed by:

$$y_i = a + b_1x_{1i} + b_2x_{2i} + b_3x_{3i} + ... + b_px_{pi} = a + \sum_{j=1}^{p} x_{ji}b_j \text{ for each } i$$

As with linear regression, the values for a and $b = (b_1, b_2, b_3, ..., b_p)$ have to be estimated, in order to find out what impact the variables $x_1, x_2, x_3, ..., x_p$ have on y. As long as p<n the value of a and $b_1, b_2, b_3, ..., b_p$ can still be estimated using the *least square method*, expressed by

$$\min_{a,b} \sum_{i=1}^{n} (y_i - a - b_1x_{1i} - b_2x_{2i} - b_3x_{3i} - ... - b_px_{pi})^2$$

where y is the response variable and x is a matrix of explanatory variables. The values of b and their significance with respect to BMD will be one of the main focuses during this project.

6.2.2 Linear multivariate regression with p>n

In this project, there have been taken biopsies from the spina iliaca of 84 voluntary women and at the same time their BMD has been measured. From each

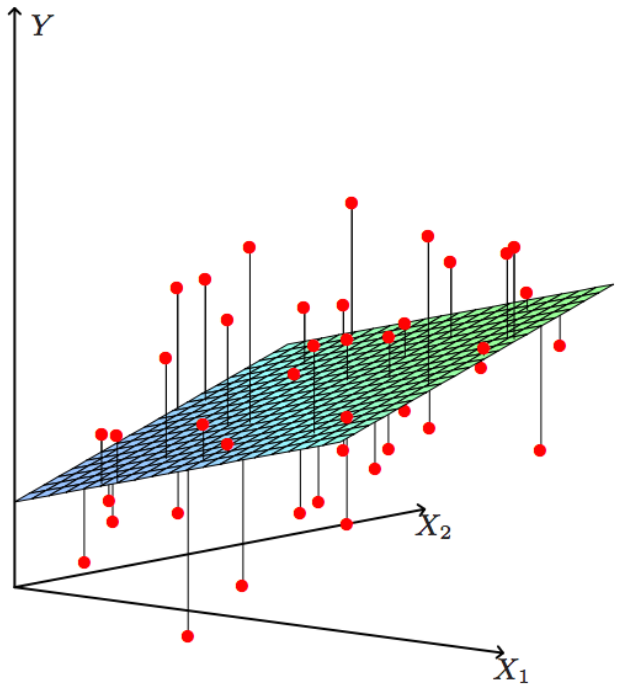


Figure 4: Linear least squares fitting with $p > 1$. We seek the linear function of $(x_1, x_2, ..., x_n)$ that minimizes the sum of squared residuals from Y.

biopsy, we get approximately p=25 000 expressions of significance. In the regression function, y_i equals BMD and $x_{1i}, x_{2i}, x_{3i}, ..., x_{pi}$ are all the genes for the i-th woman. b_{pi} is the slope for each of the x_{ip}.

We now have more independent variables, x, than samples, which means $p > n$, or $(\#genes > \#women)$. Because of this we are in danger of getting an infinite number of possible minimal values for a and b.
This is shown by this example, where $n = 2$ and $p = 3$:

$$y_1 = b_1x_{11} + b_2x_{12} + b_3x_{13}$$
$$y_2 = b_1x_{21} + b_2x_{22} + b_3x_{23}$$

We put in values to give an example:

$$3 = 10b_1 + 3b_2 + 2b_3 \; 10 = -b_1 + 2b_2 + b_3$$

Isolate b_1 on a side for the first function: $b_1 = \frac{3-3b_2-2b_3}{10}$.
Isolate b_2 on a side for the second function: $b_2 = \frac{10+b_1-b_3}{2}$.

Put the first function in the second function:

$$b_2 = \frac{10+\frac{3-3b_2-2b_3}{10}-b_3}{2} = \frac{103}{23} - \frac{12}{23}b_3.$$

As seen from the answer, the value of b_2 depends on the value of b_3 since we do not have a third equation to solve. The same will happen if we put the function for b_2 into the function of b_1:

$$b_1 = \frac{3-3(\frac{10+b_1-b_3}{2})-2b_3}{10} = \frac{36-7b_3}{17}.$$

Again we get the same problem. The value of b_1 will depend on the value of b_3. We miss a last equation for a last sample y_3 that would make us able to solve the whole set. In order to be able to estimate the value of BMD, using regression, we have somehow to select a certain amount of genes, so that we get less genes than women, $p < n$. These genes should be the genes with the strongest effect on BMD.

7 LASSO

The LASSO-function is defined as:

$$\hat{b} = \underset{b}{\operatorname{argmin}} \sum_{i=1}^{n}(y_i - b_0 - \sum_{p=1}^{g}(b_p x_{ip}))^2 + \lambda \sum_{p=1}^{g}|b_p|.$$

Here $\hat{b}$ is the result of the fit. n is the number of women, g is the number of genes, y_i is the value of the BMD for each woman, b_0 is where the line crosses the y-axis (but is not really necessary in this fit) and b_p is the coefficient for each gene x_p. The function consist of two parts. The first part of the function,

is a different way of writing the *least square*-method that we introduced earlier. It computes how the linear model $\sum_{p=1}^{\delta}(b_p x_{ip})$ approximates the measured data y_i. We call this term a measure of the fit of the model to the data. The second part, is a penalty part. Our main goal is to get the fit as small as possible, and at the same time get the penalty part as small as possible. In order to get the first part as small as possible, the formula tells us that we should have as many b's as possible to get a good approximation. But in this case, this also means that the penalty value in the second part will be large. If we decide to keep a variable x_{ip} in the linear model, with a certain coefficient $b_p > 0$, then we must pay a penalty equal to $\lambda|b_p|$

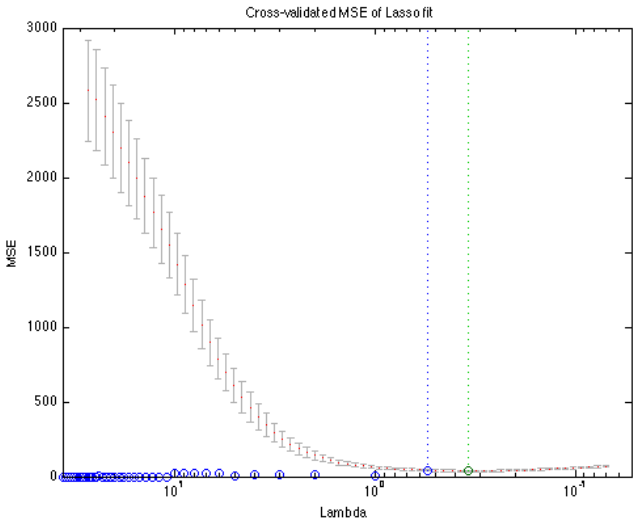
To decide which b's that will be kept, we need to determine the values of b we want to keep and in particular decide which $b_p = 0$ which means we do not want to keep them. This depends on the value of λ.

If the value of b is small, this probably means that the corresponding gene has a small impact on the outcome. If we put this b to value zero, the corresponding gene will be taken out of consideration regarding impact on BMD. If the value of b is high, the corresponding gene most likely makes a big impact on BMD. This of course will make the *least square*-part in the LASSO-formula small, but also make the penalty part large. Both solutions will increase the penalty, but it seems better with few large b's than many small b's. In the end, the goal is to reduce the number of genes as much as possible, so that the remaining genes are less than n.

In order to do this, we have to find the right value of λ. If the value of λ is large, the penalty will be large and many b's will be forced to zero. We will end up with few genes. If λ is small, fewer b's will be put to zero, and we will end up with more genes [10]. In this project, a method called cross validation will be used together with LASSO in order to find the right value of λ.

8 Cross Validation

Cross validation is a resampling method and is a much needed tool in modern statistics. It involves drawing samples from a set of observations, *the validation set* and compare these samples with fitted models from the remaining subset of observations, *the training set*. The training set uses increasing values of lambda in order to give us several fitted models. The genes found in the fitted models from the training sets are used to estimate the response variable, BMD. The estimated values of BMD is tested against the real values of BMD in the validation set. The error between the estimated value and the actual value is called the *Mean Squared error*, MSE, and we would like to find the value of λ, that gives us the smallest MSE [10]. We could have chosen to take one of the women



eksempel.png

Figure 5: The Y-axis gives us the smallest MSE, and the X-axis gives us the value of lambda that gives us the different MSE. The green dotted line shows the smallest value of lambda.

as *validation set*, and the rest of the women as the *training set*, make a fit from this, and estimated the MSE_1 and repeated this n times with a new woman as *validation set* each time. In the end we would have taken the average of all the MSE_n in order to get an estimate of the error [10]. Unfortunately, this would have taken long time if we have a large n, because the model we use is slow to fit [10]. To solve this problem, we can use *k-fold Cross validation*. Here we divide the observations into k separate groups or $folds$ of about the same size. One $fold$ will be the *Validation set*, and the remaining $k-1$ folds will be the *training set*. The MSE is computed against the first *validation set*, and is repeated k-times with a different fold as *validation set* each time. This results in k estimates of MSE, $MSE_1, MSE_2, MSE_3, ..., MSE_k$. The final estimate is computed by averaging these values, see the function [10]

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} MSE_i$$

All this is done for a fixed value of λ, and we write CV(λ). In order to choose which λ is best we take the value of λ for which the CV-error CV(λ) is smallest. For this purpose we compute CV(λ) for several values of λ in fact regularly over a grid of λ, see **Figure 5**.

9 GLMNET

In order to find the genes, we used the *lasso* package included in the Matlab R2013a version. This package is a part of the glmnet package that fits a generalized linear model via penalized maximum likelihood. It is able to regularize both by LASSO and elastic net penalty, and uses *lambda* as the regularization parameter [5]. It is authored by Jerome Friedman, Trevor Hastie, Rob Tibshirani and Noah Simon. In matlab, the function we used, looked like this:

$$[B,\ STATS] = lasso(X,Y,'CV',10).$$

The first input, X, is the matrix of all the gene expression from the 84 different women. This matrix was transposed after it was loaded into Matlab in order to get the BMD with their corresponding genes in rows instead of columns. This was done in order to make LASSO work correctly. The next input, Y, is the Bone Mineral Density-value. This input was also transposed in order to get all the women in one column. It is important to make sure that the values of the X-matrix, corresponds correctly to the Y-values. CV, says us that we will use cross validation in order to compute the MSE. *10*, tells us that we will split the 84 women in 10 different folds for the cross validation process. *B and STATS* are the outcomes . *STATS* contains information about the model fit, and this is where we find the value of λ and the indexes connected to lambda, used to find the numbers in B that not equals zero. B is a vector with the fitted coefficients. For more details see the matlab command *help lasso*.

9.1 GLMNET - example

In order to see how well LASSO works and how accurate it is on selecting the right genes, we ran a simulated test on 1000 genes. It was done by picking the 10 first genes from the probeset. The chosen genes were given a randomly selected b_p between 0.1 and 0.2 or 10 and 20 and was used to make simulated values for y for each of the 84 women. How this is done, will be explained below. In order to progress further, we had to standardize the values of the different x. We standardized the covariates x to have zero mean and variance 1, so their coefficients, *beta*, are on the same scale. In that way they can be penalized all in the same way with the same lambda.

When this was done, we generated the simulated values of y, following the model:

$$Y_i = \sum_{p=1}^{10} b_p X_{ip} + (N_i)(0, \sigma^2), \text{ where } i \text{ is each woman, } i = 1, ..., 84.$$

The first part of the model before the +-sign, is the normal way of summation to make a value Y. The second part is added in order to make the model show more variability. It is made up from a NORMAL random number distributed

B = [0.2 0.2 0.2 0.2 0.2 0.1 0.1 0.1 0.1 0.1]			
σ	False positive	False negative	Correct findings
0,1	22	2	8
0,5	24	6	4
1	21	10	0
2	0	10	0
5	8	10	0
10	0	10	0

B = [10 10 10 10 10 20 20 20 20 20]			
σ	False positive	False negative	Correct findings
0,1	9	0	10
0,5	14	0	10
1	8	0	10
2	17	0	10
5	31	0	10
10	40	0	10

Figure 6: The figure shows us the results with small and large values of the *betas* for the 10 first genes, and the influence of *sigma* the difficulty of choosing the right genes.

with mean 0 and a $\sigma^2 = 0,1$

The simulated Y_i, was used to test more genes from the probeset. We picked the first 1000 genes which also contain the 10 genes which are really relevant for y, and ran LASSO on this selection. The 1000 genes were also standardized like we did on the first 10 genes. The result from LASSO showed how well the model worked in order to pick the 10 genes we used to simulate the $Y's$. A good match shows us that LASSO is accurate and can be used in order to select which genes that influence BMD most.

The results are showed in **Figure 6**, where we change δ too.

As seen with small values of b, LASSO is able to find some of the 10 correct genes as long as the noise from *sigma* is not too big. With large b, LASSO has no problems finding the correct genes. We also see that the amount of false positive genes decreases with less noise from *sigma*. The conclusion of these findings, is that LASSO is able to choose the correct genes, as long as the beta for the gene have a big enough value. We do not know if this is the case for our BMD data, but have to assume so. If this is the case, we will find important genes. If not we might end up with false positives. A validation based on literature comparison will help us further.

10 Bone Data

The women from this study were selected from 301 non-related postmenopausal ethnic Norwegian women, aged 50-86 years. They were recruited at the outpatient clinic at Lovisenberg Deacon Hospital in Oslo. 173 were rejected because of medical reasons (underlying diseases other than OP, receiving medical treatment that might affected bone remodelling or secondary causes to OP). 23 women later decided to withdraw from the project. All the women had a normal

endocrine, biochemical, clinical and nutritional status, but differed according to the BMD of the spine, femoral neck and hip. The mean age was 64.6 years and the mean BMD was 24.2 kg/cm^2. Mean levels of vitamin K, Ca^2+, parathyroid hormone (PTH), phosphate 25(OH)vitD, carboxy-terminal telopeptide of type 1 collagen (1CTP), bone specific alkaline phosphatase and calcitonin were all within normal ranges. All groups included women that had been oestrogen users (oral contraceptives or perimenopausal treatment), but all of them had not used estrogen within 2 years prior to the sudy, except one. The remaining 100 women had trans-iliacal bone biopsies taken. 84 of these biopsies were selected for gene expression analysis. Two of the healthy women were subject for bilatetral os ileum biopsies, in order to look after molecular heterogenity. There was also taken bone biopsies from the spine (L1-L4). The same women were also measured by Total hip Dual-Energy X-ray Absorptiometry (DEXA) in order to measure BMD [11] [16].

10.1 Ethics, life style factors, BMD measurement

The Norwegian Regional Ethic Commitee gave validation and recommendations prior to the study, and all samples and procedures were according to the Law of Biobanking in Norway. All women who volunteered were given a full clinical examination, as well as laboratory analyses and DEXA of the spine (L1-L4), total hip, femoral neck and trochanter [11]. After completing questionares about several factors, one of them life style factors, the women were divided into cohorts. Since the women were Oslo-based, the cohort was seen upon as a representative for this group of women between the age of 50-86 years. All the participants, followed normal norwegian nutritional traditions: Dark bread (4 to 6 pieces/day), regular intake of milk as adults, daily intake of food containing meat or fish together with mostly potatoes, cod liver oil as vitamin D supplement in childhood, minerals (calcium and magnesium) and moderate intake of alcohol (at most, 1 to 4 glasses of wine or beer per week and seldom strong liquor). 80% took multivitamin (including vitamin A and D). The subjects were selected to three different groups according to their activity level: Group 1: active participation in physical exercise at least three times per week in addition to entertaining an active walking/hiking habit daily on the weekends. Group 2: active walking/hiking habits on the weekends but without organized exercise. Group 3: active walking related to housework, shopping, and occasionally on the weekends. From the questionaries, it was calculated that the daily intake of vitamin D and calcium was $655 \pm 80(mean \pm SD)IU/day$ and $0,73 \pm 0,45(mean \pm SD)$, respectively. The bone measurements of the L1-L4 varied with 3%. The instrumental variety of successive measurements reported on the Lunar Prodigy Instrument was $1,66\%$ for L1-L4, and $1,14\%$ for total hip. The manufacturer provided the database for determination of the Z-score [11] [16].

10.2 Bone biopsies

The bone biopsies were taken from the same place on the os ileum (2,0 cm from crista iliaca and 2 cm from spina iliaca), usually from the right side if the patient had not done hip surgery there. In the operated cases the biopsies were taken from the left side. On two of the patients, bilateral biopsies were taken. These 2 samples showed 98% similarity. The biopsies had a cylindrical diameter of 0,8 cm, and a average length of about 1,5 cm. Before they were freezed in nitrogen, connective tissue and muscle were removed. The range of weight of the biopsies was 0,35 - 0,70 g with a mean of 0,5 g. All biopsies were taken on fasting individuals in the morning. [11] [16]

10.3 Microarray analysis

With the use of GeneChip®Expression 3' Amplification One-Cycle Target Labeling Kit (Affymetrix), double stranded cDNA and biotin-labeled cRNA probes were made. According to the manufacturers instruction, the cRNA was hybridized to HG-U133 plus 2.0 chips (Affymetrix), and was then washed and stained on the GeneChip Fluidics Station 450 (Affymetrix) before they were scanned on the Affymetrix Gene Chip Scanner 3000. The quality of the RNA and probes were controlled by an Affymetrix-based test, measuring the ratio between 5' and 3' mRNAs for β-actin and GAPDH, and was found to have an agreeable level. The datas found have been submitted to the European Bioinformatics Institute (EMBL-EBI) ArrayExpress repository, ID: E-MEXP-1618 [16].

10.4 Microarray data pre-processing and evaluation

Probe sets containing more than 43 absent calls were eliminated (according to the Affymetrix MAS 5.0 software). This reduced the amount of informative probes from 54 675 to 22 185. To normalize the data, the PLIER (Probe Logarithmic Intensity Error)-algorithm was used in order to calculate relative signal values for each probe set. To show the number of present and absent calls for each probe set, there was created an Absolute Call dataset, using the MAS5 algorithm in Array Assist to discover and filter out low signal values. Correlation (Pearson) was computed between expression of each gene, and the BMD across 84 women using log transformed signal values. For each gene, zero correlation was tested against the two tailed hypothesis[16]. All this work were done by Sjur Reppe and his research group. The numbers we have been working with in this project, is a result of their preparing.

11 Results

11.1 Genes found by LASSO

We ran the LASSO function independently 100 times and ended up with unique 55 genes. These can be seen in **Figure 7**. As seen from the table, 34 of the genes are repeated more than 80 times and 22 genes were found in each of the 100 runs. There are some genes just repeated a single time, but we have chosen to still take them with us in the further evaluation.

The reason we ran LASSO 100 times, is because the genes vary for each run. It happens since we run the cross validation 10-fold. The 84 women are divided at random into 10 subsets of about 8 women each. Then lasso is run using 9 of these subsets and the last subset is used as *the validation set*. This is repeated 10 times. Every time we rerun lasso it makes a new random division into 10 subsets. This means that in each run, the CV construction is different and leads to results with some variation

11.2 Correlation

The genes found by LASSO were then correlated against the genes found in earlier research made on the gene material we had. We picked all the genes that were upregulated from the *The Molecular disease map*-study [11], since these are the genes of most importance on the gene expression. (The genes that are downregulated, are so because they are controlled to do this by the upregulated genes. This means that the genes really making an impact on the BMD will be the upregulated ones). From the paper *Eight genes are highly associated with BMD*-study [16] both negative and positive correlated genes were picked. There were about 30 genes found in other papers [12] [7] [13] [20]. Among the literature genes, 10 were not found in the probeset we used, even though we looked after aliases. They can be seen in **Figure 8**.

We ended up with about 200 genes from the literature present in our probeset and correlated them against the 55 genes found by LASSO. A part of the correlation-table can be seen in **Figure 11**. The whole table can be seen in supplemental table 1.

In **Figure 10** every gene in the literature is correlated with the 55 LASSO genes. In **Figure 11** each row is one of the 55 LASSO genes, correlated with the 200 literature genes. We see a clear tendency of confirmation of our LASSO findings.

SOST	100
ATP6V1A	100
SERINC3	100
IPO8	100
PBX3	100
PBXIP1	100
239449_at	100
DPP8	100
TFDP2	100
LRP8	100
TMEM158	100
EPM2AIP1	100
MICA	100
GTF2F2	100
RNF216	100
DUS3L	100
FKBP14	100
FAM55C	100
MAPK8	100
230672_at	100
238999_at	100
FLJ42627	100
WDR77	99
239845_at	99
238714_at	98
PIGM	98
NDUFB1	98
ATP5SL	94
SMARCA4	93
FKBP1B	85
C4orf31	85
MCF2L	85
LOC100287628	85
SP3	85
RNF213	79
TNXA /// TNXB	79
ZNF410	67
PHF3	67
KIDINS220	66
SNCA	37
BBX	37
OTUD7B	37
233646_at	37
C17orf101	21
CUX1	21
PDCD2	21
1560495_at	21
SOX1	21
IVNS1ABP	6
SERBP1	6
PBX2	4
AFFX-M27830_M_at	1
232525_at	1
DKFZp434H1419	1
MIR17HG	1

Figure 7: The table shows the 55 unique genes found after running LASSO 100 times. The numbers on the left shows how many times they were selected.

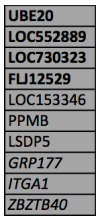


Figure 8: The table shows the genes previously known to relate to BMD not found in our probeset. The ones in bold are from [11], The normal ones are from [16] and the ones in italic are from the other articles [12] [7] [13] [20].

12 Discussion

12.1 Findings

Our findings are reported in **Figure 12**. As seen from the figure, many of the genes found by LASSO have high correlation to the literature-genes. Often we see that the LASSO genes also correlate with many of the literature genes. Starting from left in the figure, the first column on the left are the LASSO genes with highest correlations with the literature genes above 0,5. The second column shows the LASSO genes with most correlations with the literature genes above 0,35. The third column shows the LASSO genes with the highest correlation against single literature genes and the last column shows how many times the LASSO genes were selected in all of the 100 LASSO runs. This may indicate that their impact on BMD is quite strong. What is possibly interesting is that *AFFX-M27830_M_at* appeared only once in all of the 100 LASSO-runs, but is present among the literature genes. You can also see that *ATP5SL, TNXA /// TNXB, SOST, PBXIP1, SNCA, GTF2F2, DPP8, C4orf31, TFDP2 and MCF2L* all are among the top 10-11 genes in every column in **Figure 12**, which means they both have a high correlation to single literature genes, as well as they show a high correlation to many of the literature genes. They were also selected many times in all of the 100 LASSO runs. This may be an indicator that these LASSO genes have an important function in bone homoeostasis. The rest of the genes found by LASSO may also have an important role in bone homoeostasis, except for the genes *SMARCA4, 239845_at, LOC100287628, PBX2*, which did not show any correlation to any of the genes found in earlier research. These might be false positives or new discoveries, but possibly they are new findings. Validation would be necessary.

12.2 Validity

By using LASSO, we use a statistical method in order to find the genes. In order to see the validity of the LASSO-findings, we made a simulated example as seen earlier in the paper. The conclusion from that simulation (results can

Genes found in the litterature

	AFFX-M27830_M_at	SOST	RNF216	PBXIP1	MEPE	236081_at	AFFX-M27830_	WIF1	DLEU2	228528_at
SOST	0,505060189	1	-0,068852115	-0,018417301	0,955290842	0,23117987	0,508176573	0,767963005	-0,504743392	-0,413245621
AFFX-M27830_M_at	1	0,505060189	-0,248947654	-0,273769906	0,455219546	0,371742551	0,836641071	0,3978312	-0,633798465	-0,596328462
PBXIP1	-0,273769906	-0,018417301	0,205808126	1	-0,001957335	-0,334838331	-0,241589179	-0,128119211	0,081504774	0,332123953
RNF216	-0,419223284	-0,37391892	1	0,312886526	-0,329944214	-0,384482004	-0,357498424	-0,167224732	0,323304599	0,434454835
SNCA	0,408004814	0,226836275	-0,167988066	-0,323341698	0,194817248	0,898522706	0,458776902	0,062309211	-0,112047735	-0,329267532
MIR17HG	-0,817658688	-0,634789254	0,049840151	0,236787724	-0,586819933	-0,343717065	-0,737382436	-0,51121504	0,74309894	0,692035627
C4orf31	0,193100632	0,577188418	-0,156096546	0,007342737	0,58779535	-0,106831922	0,258760367	0,73161252	-0,280089092	-0,092006898
ATP5SL	-0,378310679	-0,227432441	0,279703717	0,523292364	-0,239880739	-0,420873159	-0,321666969	-0,348141276	0,211288067	0,390548862
TNXA /// TNXB	-0,346488204	-0,259132289	0,358369244	0,348929925	-0,221327175	-0,143762249	-0,299150136	-0,409033537	0,344816988	0,313220871
TFDP2	0,277640607	0,138840899	-0,142637213	-0,377933078	0,126926199	0,52845048	0,27408223	0,053551744	-0,097059211	-0,258611733
DPP8	-0,312552807	-0,239051906	0,057539533	0,310036515	-0,193256784	-0,321331063	-0,228401702	-0,328388256	0,262680559	0,539536975
MCF2L	-0,236924781	0,0465812	0,241831249	0,436594997	0,119114317	-0,322674846	-0,180805504	0,097194016	0,013370259	0,215641675
1560495_at	0,17877236	0,063789764	0,005665494	-0,152380202	0,058804114	-0,178055517	0,06673626	0,14171709	-0,110721809	-0,072069211
GTF2F2	-0,377611687	-0,27889284	0,113768878	0,323690106	-0,258603227	-0,434764437	-0,379690991	-0,269352027	0,27636848	0,321505838
EPM2AIP1	0,237325159	0,287320441	0,110110381	-0,06664128	0,260020641	0,210283226	0,181920868	0,130907465	-0,172862875	-0,082443981
ZNF410	-0,12725454	-0,209740453	0,443353468	0,224297251	-0,148077669	-0,175095809	-0,033173605	-0,136528203	0,143351966	0,086761171
FKBP1B	-0,199820538	-0,11805077	0,335340165	-0,03262022	-0,118905645	0,194292068	-0,084032874	-0,169988553	0,24628473	-0,095772736
PBX3	0,405715091	0,479012853	0,037654028	-0,0543997	0,535321356	0,173949958	0,374817748	0,36886212	-0,315038787	-0,187867445
PHF3	-0,413848716	-0,317050272	-0,031973106	0,11770449	-0,246419809	-0,419525554	-0,41543657	-0,217258344	0,307955456	0,531194357
BBX	0,433156383	0,460362979	-0,108256939	-0,137002605	0,500283613	0,117167971	0,356110837	0,466832257	-0,292091983	-0,225797462
DKFZp434H1419	-0,531024039	-0,344823111	0,127494888	0,413994158	-0,248285796	-0,236710109	-0,501112119	-0,334094387	0,278119858	0,492932479
FAM55C	-0,297342002	-0,329454443	-0,14056566	0,102008659	-0,241859692	-0,360252271	-0,292779273	-0,216320093	0,226770465	0,312006557
MAPK8	-0,360481135	-0,192953351	0,173082053	0,259479236	-0,177777548	-0,237356277	-0,26974578	-0,228570777	0,322383235	0,388076237
IVNS1ABP	-0,211823105	-0,144064696	0,32914697	0,207792918	-0,161339387	-0,246495471	-0,286224683	-0,065048228	0,138542664	0,15751755
C17orf101	-0,182292595	-0,256203129	0,232005696	0,203051582	-0,271915218	-0,096854477	-0,05731707	-0,176555522	0,195229512	0,121904797
FKBP14	0,36023913	0,394080902	-0,088869002	-0,269743411	0,365888257	0,118966289	0,353288802	0,468436242	-0,166793742	-0,223675555
IPO8	-0,233201484	-0,228639472	0,267115394	0,232445476	-0,205830892	-0,157568771	-0,124097077	-0,179716142	0,192244037	0,096390133
OTUD7B	0,065304009	0,290133768	-0,056112271	0,012406301	0,339948889	0,162609287	0,073455411	0,244799301	-0,221349624	-0,021815961
238714_at	-0,003775635	0,12468211	0,208677069	0,075250759	0,131120995	-0,026184415	0,108812967	-0,00593974	0,171619167	0,04660109
233646_at	0,446946682	0,351782044	-0,127925746	-0,227454266	0,300649429	0,217501494	0,393883773	0,236888149	-0,342097233	-0,520482735
SOX1	0,229706866	0,23850993	-0,089220608	-0,239128562	0,181237351	0,179423832	0,224491945	0,102901897	-0,16400894	-0,343739986
ATP6V1A	0,280662774	0,253779272	-0,159483367	-0,380529386	0,283489118	0,305564993	0,232897463	0,395456014	-0,264285428	-0,322470305
PIGM	0,12462239	-0,014416804	0,026978505	-0,312310512	-0,019911953	0,374162079	0,179944333	0,038288228	-0,042934516	-0,201136832
SP3	0,076830816	0,370828766	0,276924795	0,053558691	0,354747131	-0,154675243	0,089932125	0,34018679	-0,136445009	0,024294848
WDR77	-0,076927953	-0,020239654	0,011093554	0,382439114	0,016436519	-0,207008248	-0,001772028	-0,075575436	0,058476	0,201259564
DUS3L	-0,100617978	-0,26585724	0,229269448	0,166860466	-0,251138118	-0,166222601	-0,122492415	-0,299982482	0,134904072	0,086647281
FLJ42627	0,32457162	0,159663323	-0,161956695	-0,29329267	0,098667358	0,250098716	0,310932046	0,075433537	-0,181771949	-0,258275161
SERINC3	0,314295096	0,241313475	-0,217066806	-0,302737783	0,282973517	-0,011836384	0,165859083	0,245585731	-0,312404901	-0,156290035
LRP8	0,354689946	0,189100516	-0,035113392	-0,274388346	0,184843713	0,142966379	0,3263767	0,190915811	-0,292608946	-0,339840495
230672_at	0,328483072	0,253479842	-0,041452636	-0,223183054	0,195768957	0,304225415	0,302540271	0,171629697	-0,216280425	-0,460173416
NDUFB1	0,165053099	0,092304178	-0,214279076	-0,295966968	0,076339906	0,238681827	0,10983485	0,147055215	-0,094534892	-0,122888301
232525_at	0,321105813	0,359618677	-0,162420405	-0,284069327	0,314546019	0,223851994	0,299119223	0,341456696	-0,180308171	-0,256655671
238999_at	0,228328536	0,153828065	0,151123691	0,074121851	0,234866365	-0,051294637	0,12418626	0,121905814	-0,04454553	0,204962839
RNF213	0,018528737	-0,190679858	-0,127837276	-0,234161839	-0,242950886	0,011252082	-0,163323915	-0,149612106	0,087082646	0,145008263
MICA	-0,233690102	-0,19852467	0,266072422	0,143046497	-0,179532563	-0,165397045	-0,236081936	-0,271672337	0,233196011	0,016078878
SERBP1	0,1105829	-0,047888799	-0,007263785	-0,174838302	-0,078734073	0,021740863	-0,043153492	0,046698487	-0,050503478	-0,043647253
239449_at	0,322362727	0,368372417	-0,019387243	-0,216314729	0,336641519	0,306825431	0,35102407	0,289491124	-0,095920074	-0,167907006
PDCD2	0,250809956	0,090156593	-0,110457608	-0,135196236	0,08160849	0,011543695	0,115651046	0,100478739	-0,254207994	-0,109878691
CUX1	-0,148651107	-0,217456121	0,279295788	0,254055786	-0,238001862	-0,148670767	-0,158432789	-0,207010743	0,140292354	-0,035894891
TMEM158	-0,25997752	-0,230863308	-0,016712676	0,153246342	-0,198851175	-0,054810989	-0,394900367	-0,170703707	0,134522648	0,002309747
KIDINS220	0,173554478	0,17101964	0,098660955	-0,231371617	0,261560407	0,135882557	0,137863474	0,171877276	-0,10578806	-0,115725824
239845_at	0,000670605	-0,174439361	0,096740576	-0,011670138	-0,18087947	0,003528482	-0,028285581	-0,261012769	0,080556343	0,093552287
SMARCA4	-0,230529769	-0,340204436	-0,072292821	-0,0020351	-0,276039778	0,019099587	-0,140635564	-0,307909043	0,194063299	0,10784804
PBX2	0,076621447	-0,047311145	-0,041431957	-0,215837495	-0,059803361	0,049023325	-0,019111681	-0,038949189	0,144965876	0,002868841
LOC100287628	-0,037086209	0,074383504	-0,01668299	-0,229444502	0,053357218	0,12715671	-0,050034707	0,162881187	-0,040236682	0,030395375

Figure 9: Correlation between the 10 most correlated genes found in the literature and the genes found by LASSO, with the most correlation on the top. Correlation >0,35 is marked with red and correlation >0,5 is marked yellow. Genes found in both the literature and by LASSO have correlation = 1 and is marked green. The literature genes in bold are from [11], the normal ones are from [16] and the ones in italic are from the other articles [12] [7] [13] [20].

Genes found in literature												
AFFX-M27830_M_at	SNCA	PBX3	BBX	LRP8	FKBP14	233646_at	SOST	AFFX-M27830_M_at
AFFX-M27830_5_at	SNCA	PBX3	239449_at	BBX	FKBP14	233646_at	SOST	AFFX-M27830_M_at
SOST	PBX3	239449_at	BBX	C4orf31	232525_at	FKBP14	233646_at	SP3	SOST	AFFX-M27830_M_at
MMP13	BBX	C4orf31	sost	AFFX-M27830_M_at
DKK1	PBX3	239449_at	BBX	C4orf31	sost	AFFX-M27830_M_at
TMED4	SERINC3	SERBP1	FKBP14	NDUFB1	FLJ42627
MEPE	PBX3	BBX	C4orf31	FKBP14	SP3	SOST	AFFX-M27830_M_at
SOX4	BBX	PDCD2
PTMS	DUS3L
AP2A1	CUX1
BCL2	SNCA	TFDP2
IQCE	ZNF410
244788_at	MCF2L	SP3	SOST
EXOC7	ZNF410	DUS3L
MAZ	ZNF410	FKBP1B
SAMD4B	EPM2AIP1
ZNF277	C17orf101	FAM55C	RNF216
EXOSC4	ZNF410	MICA
241893_at	BBX	LRP8	EPM2AIP1	1560495_at
EMP3	LRP8
LARP1	ZNF410
AP1G2	IVNS1ABP	LRP8	EPM2AIP1
SNORD104	EPM2AIP1
SYK	IPO8	IVNS1ABP	PBX3	LRP8
RNASEH2C	PBX3	LRP8	230672_at
ASF1B	ZNF410	FKBP1B	PIGM
ACIN1	ZNF410	DUS3L
TAF6	SOX1
DGKA	PBX3	EPM2AIP1	DUS3L
BOLA1	PBX3	LRP8
228639_at	OTUD7B
MON1B	RNF213
NUP210	SNCA	TFDP2	FKBP1B	PIGM
LAS1L	ZNF410	FKBP1B	DUS3L
225722_at	PBX3	233646_at
GYPC	SNCA	TFDP2	FKBP1B
GTF2F1	PBX3	ZNF410
RAD9A	232525_at
ZDHHC3	KIDINS220
SF3A1	SNCA	FKBP1B
TH1L	PBX3
GINS3	SNCA	ATP6V1A	PBX3	238714_at
FAM91A2	SERINC3
RCC2	PBX3	LRP8	230672_at	AFFX-M27830_M_at
PRELID1	PBX3
SYNGR1	SNCA	TFDP2	FKBP1B	LRP8	230672_at
CXXC1	230672_at	233646_at
BLVRB	SNCA	TFDP2	FKBP1B	PIGM
LASS2	PBX3
PPME1	SNCA	TFDP2	FKBP1B
BCL2L12	230672_at
SEC61B	ATP6V1A	PIGM	AFFX-M27830_M_at
PRDX2	SNCA	TFDP2	FKBP1B	230672_at
FRMD4A	SNCA	TFDP2	FLJ42627
WIF1	ATP6V1A	PBX3	BBX	C4orf31	FKBP14	SOST	AFFX-M27830_M_at
EIF2AK1	SNCA	ATP6V1A	TFDP2	FKBP1B
236081_at	SNCA	TFDP2	PIGM	AFFX-M27830_M_at
RPS10	PBX3	LRP8	EPM2AIP1	PDCD2
241091_at	SERINC3	LRP8
PIK3R2	SNCA	TFDP2	FKBP1B
S100A6	ZNF410
225902_at	TNXA /// TNXB
VWF	ATP5SL	TNXA /// TNXB	GTF2F2	MCF2L	PBXIP1	RNF216
LOC643733	MAPK8	PBXIP1
AGL	ATP5SL	DPP8	TNXA /// TNXB	GTF2F2	FAM55C	MCF2L	DKFZp434H1419	MAPK8	MIR17HG	PBXIP1	RNF216
226392_at	C17orf101	MAPK8
237157_at	C4orf31	MAPK8
240499_at	SP3
RNF20	IPO8	C17orf101	PBXIP1
ACSL3	ATP5SL	MCF2L
BTBD1	ATP5SL	DPP8	TNXA /// TNXB	GTF2F2	FAM55C	MCF2L	DKFZp434H1419	MIR17HG	PBXIP1	RNF216
ITIH5	WDR77	MCF2L	PBXIP1
CLASP2	DPP8	PHF3
PRKCQ	ATP5SL	DPP8	TNXA /// TNXB	GTF2F2	MCF2L	DKFZp434H1419	PBXIP1	RNF216
227107_at	ATP5SL	TNXA /// TNXB	TMEM158	GTF2F2	PBXIP1
UBE3A	ATP5SL	TNXA /// TNXB	PBXIP1
PPARA	ATP5SL	TNXA /// TNXB	GTF2F2	238999_at	PBXIP1
UBLCP1	C17orf101
LRP12	GTF2F2
CCBL2	238714_at
SCN1B	ATP5SL	TNXA /// TNXB	DKFZp434H1419	MIR17HG
KLF10	GTF2F2	RNF216
PURB	PBX3	ZNF410	238714_at
GCSH	ATP5SL	DPP8	TNXA /// TNXB	PHF3	GTF2F2	FAM55C	MCF2L	DKFZp434H1419	PBXIP1	RNF216
MSRB2	ATP5SL	TNXA /// TNXB	PBXIP1
MRS2	ATP5SL	DPP8	TNXA /// TNXB	GTF2F2	DKFZp434H1419	MAPK8	MIR17HG	RNF216
ITFG1	IPO8	C17orf101	RNF216
C19orf12	ATP5SL	DPP8	TNXA /// TNXB	WDR77	GTF2F2	MCF2L	OTUD7B	PBXIP1
GRSF1	ATP5SL	DPP8	PBXIP1
FAM134B	ATP5SL	DPP8	TNXA /// TNXB	GTF2F2	FAM55C	DKFZp434H1419	MAPK8	PBXIP1	RNF216
KPNA4	ATP5SL	DPP8	TNXA /// TNXB	GTF2F2	FAM55C	MCF2L	PBXIP1
RANBP6	IPO8	PBXIP1
LAPTM4B	ATP5SL	TNXA /// TNXB	GTF2F2	MCF2L	PBXIP1	RNF216
FAM35A	IPO8	DPP8	PHF3	FAM55C
226297_at	ATP5SL	TNXA /// TNXB	GTF2F2	MCF2L	MAPK8	MIR17HG	PBXIP1	RNF216
RNF216	TNXA /// TNXB	ZNF410	RNF216
244511_at	ATP5SL	DPP8	TNXA /// TNXB	GTF2F2	MCF2L	DKFZp434H1419	PBXIP1	RNF216
ATPAF1	ATP5SL	DPP8	TNXA /// TNXB	GTF2F2	PBXIP1
PHTF2	ATP5SL	DPP8	TNXA /// TNXB	GTF2F2	DKFZp434H1419	MIR17HG	PBXIP1
PAIP1	TNXA /// TNXB	FAM55C
KIAA0368	ATP5SL	DPP8	TNXA /// TNXB	PHF3	GTF2F2	FAM55C	DKFZp434H1419	PBXIP1	RNF216
DLEU2	MIR17HG
USP47	ATP5SL	TNXA /// TNXB	PBXIP1	RNF216
LAMB1	MCF2L	MAPK8	PBXIP1
CRIM1	PHF3	GTF2F2	RNF216
GNG11	MCF2L
KIF1B	ATP5SL	DPP8	TNXA /// TNXB	GTF2F2	DKFZp434H1419	PBXIP1	RNF216
DNAJC19	238999_at
SLC16A1	DPP8	TNXA /// TNXB	GTF2F2	FAM55C	DKFZp434H1419	MAPK8
SLC44A1	IVNS1ABP	1560495_at
228464_at	DPP8	PHF3	MIR17HG
PTP4A1	ATP5SL	DPP8	TNXA /// TNXB	GTF2F2	FAM55C	MCF2L	DKFZp434H1419	MAPK8	PBXIP1	RNF216
USP25	DPP8	FAM55C	238999_at
REEP1	ATP5SL	DPP8	TNXA /// TNXB	GTF2F2	FAM55C	MAPK8	PBXIP1
WWP1	OTUD7B
FKBP3	ATP5SL	DPP8	TNXA /// TNXB	GTF2F2	FAM55C	DKFZp434H1419	PBXIP1	RNF216
PHKB	PHF3	FAM55C
RTN4	ATP5SL	DPP8	TNXA /// TNXB	GTF2F2	FAM55C	MCF2L	DKFZp434H1419	PBXIP1
TSNAX	IPO8	C17orf101	RNF216
PTGER4	C17orf101	MIR17HG
PBXIP1	ATP5SL	WDR77	MCF2L	DKFZp434H1419	PBXIP1
AKTIP	ATP5SL	DPP8	GTF2F2	FAM55C	DKFZp434H1419	MAPK8
KTN1	ATP5SL	TNXA /// TNXB	GTF2F2	MCF2L	DKFZp434H1419	PBXIP1	RNF216
228528_at	ATP5SL	DPP8	PHF3	DKFZp434H1419	MAPK8	MIR17HG	RNF216
SBDS	ATP5SL	DPP8	TNXA /// TNXB	PHF3	GTF2F2	FAM55C	MCF2L	DKFZp434H1419	OTUD7B	PBXIP1	RNF216
230127_at	MIR17HG	RNF216
HSPA9	ATP5SL	238999_at
SELT	SERINC3	238999_at
SLC25A30	ATP5SL	DPP8	TNXA /// TNXB	C17orf101	GTF2F2	FAM55C	MCF2L	DKFZp434H1419	MAPK8	MIR17HG	PBXIP1	RNF216
LOH3CR2A	ATP5SL	DPP8	TNXA /// TNXB	GTF2F2	FAM55C	MCF2L	DKFZp434H1419	RNF216
PPP3CB	ATP5SL	DPP8	TNXA /// TNXB	GTF2F2	MCF2L	DKFZp434H1419	MAPK8	PBXIP1
UBLCP1	C17orf101
RBM18	DPP8	PHF3	FAM55C	RNF216
NIPSNAP3B	ATP5SL	DPP8	TNXA /// TNXB	PHF3	GTF2F2	FAM55C	MCF2L	DKFZp434H1419	MAPK8	RNF216
SLC38A2	ATP5SL	DPP8	DKFZp434H1419	MAPK8	PBXIP1
1558801_at	ATP5SL	DPP8	GTF2F2	PBXIP1	RNF216
ARHGAP18	FKBP1B
MRPL19	GTF2F2	MAPK8
TDRD3	MCF2L
HMGN3	MAPK8	PBXIP1
OSBPL1A	ATP5SL	DPP8	TNXA /// TNXB	PHF3	GTF2F2	FAM55C	MCF2L	DKFZp434H1419	MAPK8	MIR17HG	PBXIP1	RNF216
YPEL2	ATP5SL	DPP8	TNXA /// TNXB	MCF2L	MAPK8	PBXIP1
AP1S2	PHF3
COPS4	ATP5SL	DPP8	TNXA /// TNXB	C17orf101	GTF2F2	FAM55C	MCF2L	PBXIP1
CLIP4	ATP5SL	DPP8	TNXA /// TNXB	PHF3	GTF2F2	FAM55C	MCF2L	DKFZp434H1419	MAPK8	MIR17HG	PBXIP1	RNF216
ZYG11B	ATP5SL	DPP8	TNXA /// TNXB	GTF2F2	FAM55C	MCF2L	DKFZp434H1419	PBXIP1	RNF216
C17orf91	DPP8	TNXA /// TNXB	DKFZp434H1419	MIR17HG
VDR	PBX3	BBX	SOST
ESR2	DUS3L
LRP4	C4orf31	MCF2L	OTUD7B	SOST
RANKL	233646_at	SOX1
COL1A1	C4orf31
SPP1	C4orf31	SOST
SP7	C4orf31	SOST
SOX6	SNCA	ATP6V1A	TFDP2	PIGM	232525_at	FLJ42627	AFFX-M27830_M_at
ADAMTS18	FLJ42627
TGFBR3	MCF2L	OTUD7B
ARHGEF3	C17orf101
CYP19A1	233646_at
DBP	DPP8	TNXA /// TNXB	FAM55C	DKFZp434H1419	MAPK8	RNF216
FLNB	C4orf31	MCF2L	SP3	SOST
FOXC2	C4orf31	MCF2L	SOST

Figure 10: Each row shows one of the literature genes and which LASSO-genes they correlate with. The colours show grade of correlation. Correlation >0,35 = red, correlation >0,5 = yellow and correlation = 1 = green.

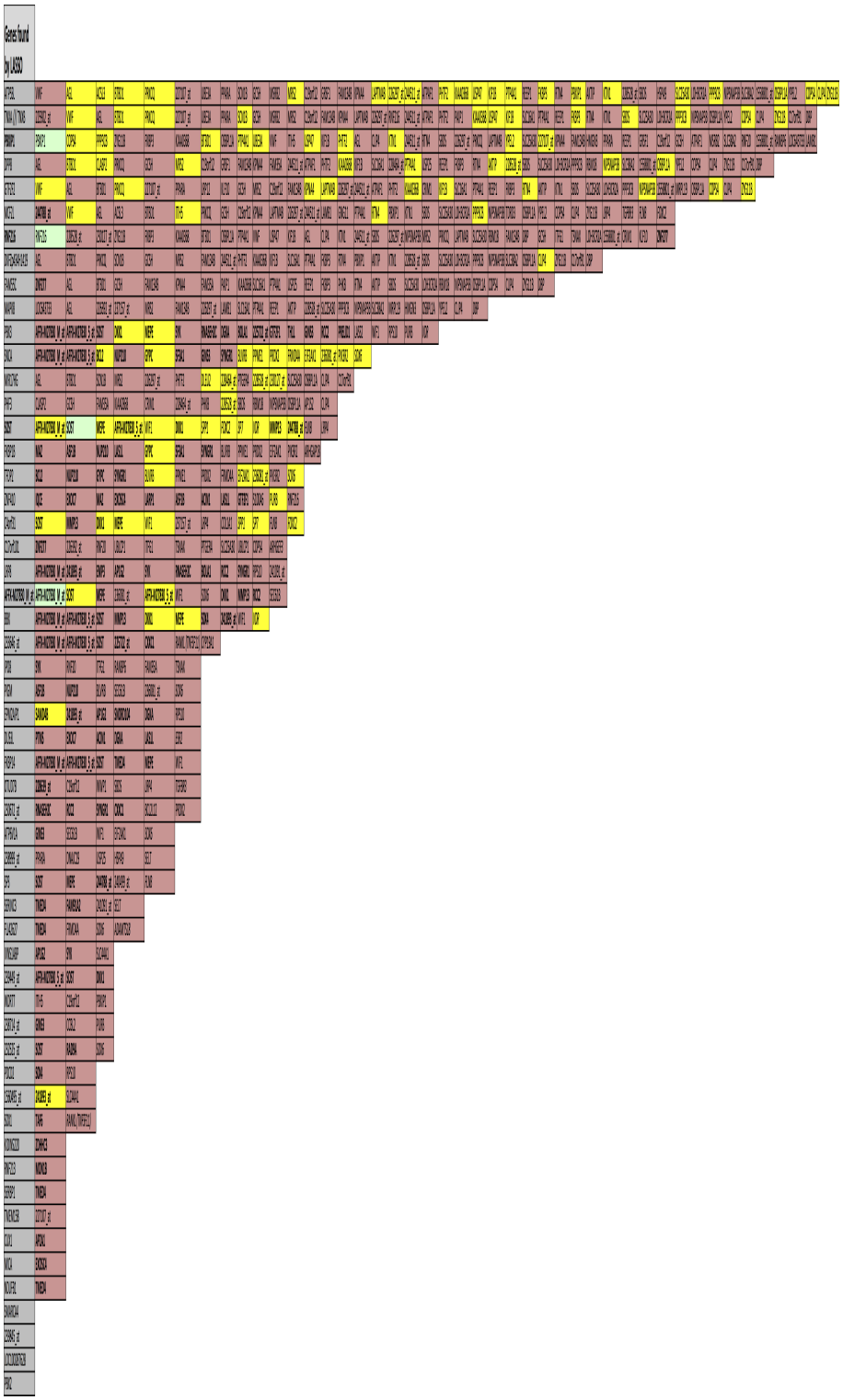


Figure 11: Each row shows one of the LASSO-genes and which literature-genes they correlate with. The colours show grade of correlation. Correlation >0,35 = red, correlation >0,5 = yellow and correlation = 1 = green.

Most corr > 0,5	Number	Most corr > 0,35	Number	Highest corr	Correation	Most repeats LASSO	Number
ATP5SL	22	ATP5SL	45	**SOST**	1	**SOST**	100
TNXA /// TNXB	12	TNXA /// TNXB	43	**AFFX-M27830_M_at**	1	ATP6V1A	100
SOST	12	**PBXIP1**	43	**PBXIP1**	1	SERINC3	100
PBXIP1	11	DPP8	40	**RNF216**	1	IPO8	100
SNCA	10	GTF2F2	39	SNCA	0,898522706	PBX3	100
GTF2F2	10	MCF2L	34	MIR17HG	0,743099894	**PBXIP1**	100
DPP8	9	**RNF216**	34	C4orf31	0,73161252	239449_at	100
C4orf31	7	DKFZp434H1419	30	ATP5SL	0,625462549	DPP8	100
TFDP2	4	FAM55C	27	TNXA /// TNXB	0,595967383	TFDP2	100
MCF2L	4	MAPK8	23	TFDP2	0,593008327	LRP8	100
MIR17HG	4	PBX3	20	DPP8	0,579387183	TMEM158	100
BBX	3	SNCA	16	MCF2L	0,576237585	EPM2AIP1	100
AFFX-M27830_M_at	3	MIR17HG	15	1560495_at	0,569622147	MICA	100
PBX3	2	PHF3	14	GTF2F2	0,549386738	GTF2F2	100
ZNF410	1	**SOST**	14	EPM2AIP1	0,547700491	**RNF216**	100
FKBP1B	1	FKBP1B	13	ZNF410	0,539710376	DUS3L	100
PHF3	1	TFDP2	12	FKBP1B	0,536758013	FKBP14	100
EPM2AIP1	1	ZNF410	12	PBX3	0,535321356	FAM55C	100
DKFZp434H1419	1	C4orf31	12	PHF3	0,531194357	MAPK8	100
1560495_at	1	C17orf101	11	BBX	0,514226658	230672_at	100
RNF216	1	LRP8	11	DKFZp434H1419	0,504653725	238999_at	100
ATP6V1A	0	**AFFX-M27830_M_at**	11	FAM55C	0,498363666	FLJ42627	100
SERINC3	0	BBX	10	MAPK8	0,496167197	WDR77	99
KIDINS220	0	233646_at	7	IVNS1ABP	0,480893571	239845_at	99
IPO8	0	IPO8	6	C17orf101	0,475020133	238714_at	98
IVNS1ABP	0	PIGM	6	FKBP14	0,468436242	PIGM	98
239449_at	0	EPM2AIP1	6	IPO8	0,466245746	NDUFB1	98
RNF213	0	DUS3L	6	OTUD7B	0,464595343	ATP5SL	94
WDR77	0	FKBP14	6	238714_at	0,464212546	SMARCA4	93
238714_at	0	OTUD7B	6	233646_at	0,446946682	FKBP1B	85
C17orf101	0	230672_at	6	SOX1	0,443653626	C4orf31	85
LRP8	0	ATP6V1A	5	ATP6V1A	0,438403489	MCF2L	85
PIGM	0	238999_at	5	PIGM	0,431916675	LOC100287628	85
SERBP1	0	SP3	5	SP3	0,425396372	SP3	85
TMEM158	0	SERINC3	4	WDR77	0,425123924	RNF213	79
CUX1	0	FLJ42627	4	DUS3L	0,424475412	TNXA /// TNXB	79
MICA	0	IVNS1ABP	3	FLJ42627	0,423234714	ZNF410	67
232525_at	0	239449_at	3	SERINC3	0,421119551	PHF3	67
PDCD2	0	WDR77	3	LRP8	0,41862584	KIDINS220	66
DUS3L	0	238714_at	3	230672_at	0,412367739	SNCA	37
FKBP14	0	232525_at	3	NDUFB1	0,410279799	BBX	37
FAM55C	0	PDCD2	2	232525_at	0,406244461	OTUD7B	37
NDUFB1	0	1560495_at	2	238999_at	0,403087463	233646_at	37
OTUD7B	0	SOX1	2	RNF213	0,38248637	C17orf101	21
MAPK8	0	KIDINS220	1	MICA	0,376062466	CUX1	21
230672_at	0	RNF213	1	SERBP1	0,372866019	PDCD2	21
233646_at	0	SERBP1	1	239449_at	0,368372417	1560495_at	21
238999_at	0	TMEM158	1	PDCD2	0,367417567	SOX1	21
FLJ42627	0	CUX1	1	CUX1	0,360454052	IVNS1ABP	6
SOX1	0	MICA	1	TMEM158	0,357928665	SERBP1	6
SP3	0	NDUFB1	1	KIDINS220	0,351055606	PBX2	4
SMARCA4	0	SMARCA4	0	239845_at	0,303649572	**AFFX-M27830_M_at**	1
239845_at	0	239845_at	0	SMARCA4	0,299090397	232525_at	1
LOC100287628	0	LOC100287628	0	PBX2	0,279669187	DKFZp434H1419	1
PBX2	0	PBX2	0	LOC100287628	0,245155905	MIR17HG	1

Figure 12: Starting from left in the figure, the first column on the left are the LASSO genes with highest correlations above 0,5 with the literature genes. The second column shows the LASSO genes with most correlations above 0,35 with the literature genes. The third column shows the LASSO genes with the highest correlation against single literature genes and the last column shows how many times the LASSO genes were selected in all of the 100 LASSO runs.

be seen in **figure 6**), was that as long as the *Beta*-values were big enough, we should be able to find all the correct genes, but we are also in danger of finding false positives when the noise is strong. Beta values say how much the actually important genes affect BMD, in absolute value. We do not know this, but on the other hand small effects (which we would not find) might be important.

We find that 51 of the LASSO genes correlates more than 0,35 with the expression of genes found in the literature. 21 of the genes correlates more than 0,5. This indicates that our 51 genes are true findings. We also checked if any of the LASSO had aliases that were the same as the literature-genes, which it was not. The conclusions of our findings is that most of the LASSO-findings are reliable, as it is correlated to genes already known being associated to bone mechanisms.

12.3 Limitations

The genes used in this experiment, are from biopsies taken from the hip. The Z-values we used for BMD, were only adjusted for age. We might have got some bias due to confounding factors such as BMI which we have not investigated this time. There might be other influencing factors that make some bias that we do not know about. The cohort was well designed, taking into account influencing factors such as nutrition, activity and daily intake of vitamin D and Calcium [11] [16]. On the other hand, the Z-values are not adjusted for blood values of PTH, Calcium or 25(OH)vitD. Since the women were Oslo-based, the cohort may be seen as representative for this group of women between the age of 50-86 years. We can not generalize the findings much further.

12.4 Functional assessment

The biological function of a selection of the 20 highest ranked genes from **Figure 12** are shown in **Figure 13**. The genes shown are those in which we found most information about in the literature.

Of the function and pathways seen, *SOST* seems to be the one of most interest. It is a gene known for being a negative regulator of bone-growth. It is also involved in the *Wnt-pathway*, which is central in bone turnover. Another interesting gene is *MCF2L*, which seems to have some interaction with the gene *RHOA*.

12.5 Conclusion

The LASSO is a method for finding genes involved in biological processes. It seems to be a valid method for selecting the right genes. We found 51 genes highly associated with BMD and mostly correlated to genes already found in the literature for having an impact on BMD. Among these, we found one gene, $SOST$, that for sure is known to have impact on BMD. Another gene, $MCF2L$, seems to have some cooperation with $RHOA$. 4 genes found by LASSO were

	Final text
SOST	Inhibit the Wnt signalingpathway and work as a negative regulator of bone growth.
PBXIP1	Regulation of pre-B-cell leukemia transcription factors (BPXs). Blocks the transcriptional activity of E2A-PBX1 through inhibition of the binding of PBX1-HOX complex to DNA. Also binds estrogen receptor-alpha (ESR1) to microtubules and by this allows them to influence estrogen receptors-alpha signaling.
RNF216	Has a Isoform 1 which acts as an E3 ubiquitin ligase. It accepts ubiquitin from specific E2 ubiquitin-conjugating enzymes, and then transfers it to substrates which is degratated by the proteasome. It also promotes degradation of TRAF3, TLR4 and TLR9, which take part in the regulation of antiviral responses, negative activation of NF-kappa-B, IRF3 activation and IFNB production. The 3/ZIN Isoform, down-regulates TNF and IL-1 mediated activation of NF-kappa-B. It promotes TNF and RIP mediated cell death.
SNCA	The gene may be involved in the regulation of dopamine release and transport and this may also explain why it is included in a Parkinson disease-pathway. Induces fibrillization of microtubule-associated protein tau, which leads to a decreased caspase-3 activation due to reduced neuronal responsiveness to various apoptotic stimuli . Its mutant phenotype affects the Skeleteon.
TNXA /// TNXB	**TNXA:** is homologous to XB and is included in YA and overlapping CYP21P, both in opposite orientation. **TNXB:** **M**ediates interactions between cells and the extracellular matrix. It inhibits cell migration by working as a substrate-adhesion molecule, induces collagen fibril formation and may play a role in supporting epithelial tumors growth.
TFDP2	Stimulate E2F-dependent transcription. Binds DNA cooperatively with E2F family members transcription factors through the E2 recognition site, 5'-TTTC[CG]CGC-3', which is found in the promoter region of several genes whose products are involved in cell cycle regulation or in DNA replication. It makes a DP2/E2F complex function which contro cell-cycle progression from G1 to S phase. Theres is also a E2F1/DP complex that appears to mediate both cell proliferation and apoptosis. Is also found to be in the PDGF signaling pathway.
DPP8	Is a dipeptidyl peptidase and cuts off N-terminal dipeptides from proteins having a Pro or Ala residue at position 2. May play a role in activation of T-cells and immune function.
GTF2F2	Is a general transcription factor that binds to RNA polymerase II and recruits it to the initiation complex. Helps in transcription elongation. Also shows ATP-dependent DNA-helicase activity. Is in the pathway for transcription regulation by bZIP transcription factor.
MCF2L	Guanine nucleotide exchange factor that very likely links pathways that signal through RAC1, RHOA and CDC42. Works as a catalysator for guanine nucleotide exchange on RHOA and CDC42 and interacts specifically with the GTP-bound form of RAC1. This may indicate that it functions as an effector of RAC1. May also take part in axonal transport in the brain. Becomes activated and highly tumorigenic by truncation of the N-terminus. The isoform 5 of the gene activates CDC42.

Figure 13: Functions and pathways for some of the top 10-20 ranked genes from **Figure 12** in which we found most information about [8] [15]

.

previously known as regulating genes. Our study strengthens the probability that they really have an impact on BMD. There were 4 new genes, not correlated to any gene known previously to play a role in bone mechanisms, and these might be new discoveries or false positives. To find this out, we need to do further investigations. The rest of the LASSO genes seems to be highly associated to BMD.

References

[1] O.O. Aalen. *Statistiske metoder i medisin og helsefag: Odd O. Aalen (red.).* Gyldendal akademisk, 2006.

[2] C Champe, Richard A Harvey, et al. *Lippincott's IIIustrated Reviews of Biochemistry*. Lippincott Williams and Wilkins A Wolters Kluwer Company, USA, 2008.

[3] Universidade Federal do Rio Grande do Sul. Osteon. `http://www.ufrgs.br/imunovet/molecular_immunology/osteon.jpg`.

[4] Bruce Ettinger, Dennis M Black, Michael C Nevitt, Amy Chen Rundle, Jane A Cauley, Steven R Cummings, and Harry K Genant. Contribution of vertebral deformities to chronic back pain and disability. *Journal of Bone and Mineral Research*, 7(4):449–456, 1992.

[5] Trevor Hastie and Junyang Qian. Glmnet vignette.

[6] Per Holck. Osteologi en innfoering.

[7] Takayuki Hosoi. Genetic aspects of osteoporosis. *Journal of bone and mineral metabolism*, 28(6):601–607, 2010.

[8] Weizmann institute of science. Genecards, the human gene compendium. `http://www.genecards.org`.

[9] St. Rosemary Education Institution. Howship lacunae. `http://schoolworkhelper.net/wp-content/uploads/2013/02/Howship-lacunae-slide-labelled-histology.jpg`.

[10] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*. Springer, 2013.

[11] Rune Jemtland, Marit Holden, Sjur Reppe, Ole K Olstad, Finn P Reinholt, Vigdis T Gautvik, Hilde Refvem, Arnoldo Frigessi, Brian Houston, and Kaare M Gautvik. Molecular disease map of bone characterizing the postmenopausal osteoporosis phenotype. *Journal of bone and mineral research*, 26(8):1793–1801, 2011.

[12] Wen-Feng Li, Shu-Xun Hou, Bin Yu, Meng-Meng Li, Claude Férec, and Jian-Min Chen. Genetics of osteoporosis: accelerating pace in gene identification and validation. *Human genetics*, 127(3):249–285, 2010.

[13] Braxton D Mitchell and Laura M Yerges-Armstrong. The genetics of bone loss: challenges and prospects. *The Journal of Clinical Endocrinology & Metabolism*, 96(5):1258–1268, 2011.

[14] Robert Nussbaum, Roderick R McInnes, and Huntington F Willard. *Thompson & Thompson genetics in medicine*. Elsevier Health Sciences, 2007.

[15] U.S. National Library of Medicine. National center for biotechnology information. `http://www.ncbi.nlm.nih.gov/gene`.

[16] Sjur Reppe, Hilde Refvem, Vigdis T Gautvik, Ole K Olstad, Per I Høvring, Finn P Reinholt, Marit Holden, Arnoldo Frigessi, Rune Jemtland, and Kaare M Gautvik. Eight genes are highly associated with bmd variation in postmenopausal caucasian women. *Bone*, 46(3):604–612, 2010.

[17] Stanley Leonard Robbins, Vinay Kumar, Abul K Abbas, and Jon C Aster. *Robbins basic pathology*. Elsevier Health Sciences, 2012.

[18] Tom Strachan and Andrew Read. *Human molecular genetics*. Garland Science, 2010.

[19] Barbara Young, Phillip Woodford, and Geraldine O'Dowd. *Wheater's functional histology: a text and colour atlas*. Elsevier Health Sciences, 2013.

[20] Hou-Feng Zheng, Timothy D Spector, and J Brent Richards. Insights into the genetics of osteoporosis from recent genome-wide association studies. *Expert reviews in molecular medicine*, 13:e28, 2011.